De-unifying a Digital Library

by Arthur Sale

Abstract

The University of Tasmania decided to explore using a unified digital library for all its research output: journal articles, conference papers, higher degree theses, and other types. This decision is in advance of the state of the Australian national indexing systems. The digital library also uses OAI-PMH protocols for harvesting, which one of the national repositories does not as yet. The paper describes the context, reasons for the University's decision, consequences and outcomes, and the development of software to talk to the Australian Digital Theses Program.

Contents

Context

Australian national initiatives

Australia has a number of national digital library initiatives. Two are relevant to this project: *Australian Digital Theses Program* (ADTP, 2004a) and *Australian Research Repositories Open to the World* (ARROW, 2004).

ADTP started in 2000, and provides a national metadata repository for research higher degree theses (for example PhD theses) at Australian universities. The full-text of the theses are actually held in university repositories, and the national repository harvests the metadata daily through the Internet to provide the central search service. Viewers wanting to examine a particular thesis are referred to the local repository. To date 23 Australian universities are active contributors, out of a total of 39. Neither the central repository nor any of the local repositories, except the University of Tasmania, are compliant with the *Open Access Initiative Protocol for Metadata Harvesting* (OAI-PMH, 2004). However this is planned for the near future

ARROW is in start-up mode. One of ARROW's two functions is to provide a national search interface to all Australian research output, including journal articles, conference papers, theses, etc. From the start, ARROW is designed

to be OAI-PMH compliant, to harvest from OAI-PMH-compliant local repositories, and to provide for OAI-PMH harvesting of its metadata by higher level service providers. Seven Australian universities operate OAI-PMH compliant research repositories at present and are represented in the interim ARROW search engine, but this is expected to grow rapidly once ARROW is in full operation and provides support for its preferred repository package *Fedora* with a *VTLS* overlay. ARROW will harvest from ADTP as this forms part of the Australian research output.

The Australian Government announced in October 2003 that ARROW and ADTP would be two of four recipients in a $A12M national digital library initiative.

University of Tasmania

The *University of Tasmania* (UTas) is one of Australia's top ten research universities and the only one in Australia's island state. In early 2004, the University explored mounting *all* its research output (research papers *and* theses) in a single digital library compliant to OAI-PMH standards. This allows the University's research output to be harvested direct by specialized OAI search engines such as ARC and OAIster, and also by general search engines such as Yahoo and Google.

The University will then not be entirely dependent on either ARROW or ADTP for exposure of its research, and can maintain its reputation for ICT innovation. The single unified repository would be simple to maintain and operate. All research output would have the same exposure and be searchable through the same interfaces. Long term archiving would also be simplified since it would be dealt with in a unified way.

Figure 1 shows the harvesting and search relationships between the entities described earlier.

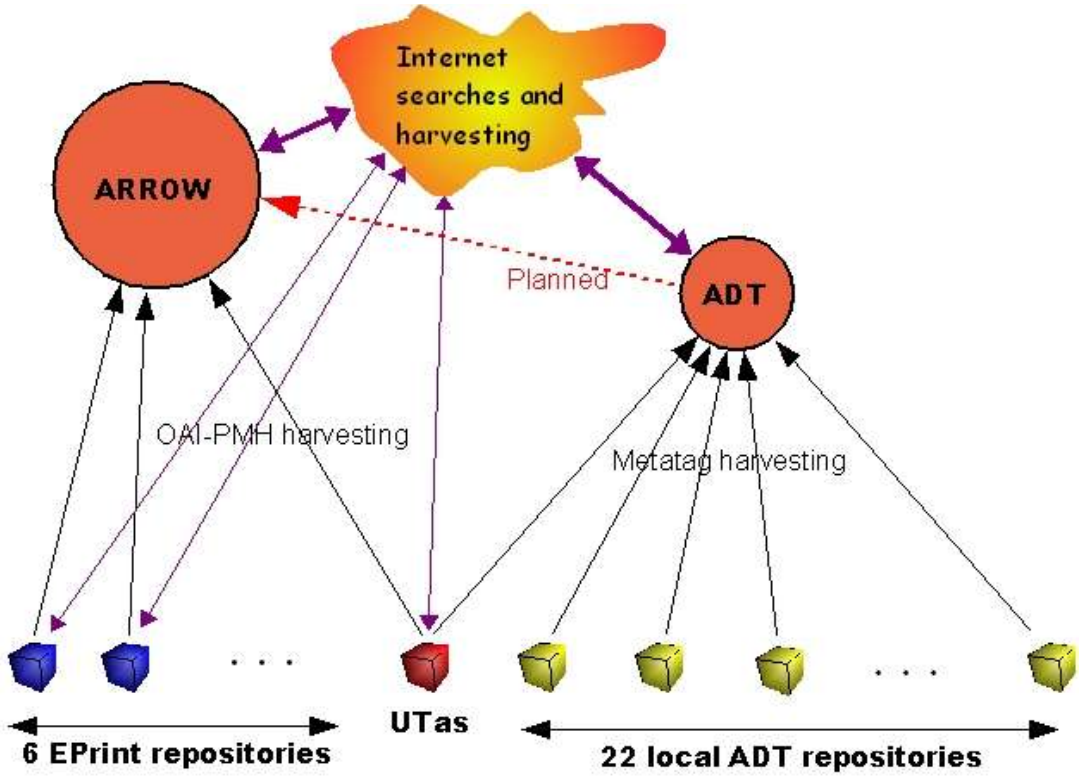

Figure 1 – Relationship between Australian repositories

EPrints 2.2.1 software was chosen as a prototype, and the UTas server went live in May 2004 at http://eprints.comp.utas.edu.au/ (UTasER, 2004). *EPrints* (2004) is free open source software, and uses also free open-source packages Apache, Perl, and mySQL, running on top of a Unix OS. Over 60% of OAI repositories in the world run on EPrints. The decision has been confirmed as a good one: there have been absolutely no bugs or crashes since the system went operational, and uploading documents has been easy. Minimal ICT support has been needed.

In the longer term, the UTas repository may be migrated to *Fedora* (2004), chosen by ARROW as its recommended local repository. This should be a relatively easy transition, as both are OAI-PMH compliant and Fedora has a bulk import facility. Fedora is also open source, but the ARROW recommendation will have a commercial VTLS overlay (2004). Fedora is a more sophisticated package than EPrints and provides for a variety of digital objects. The issues raised in this paper will need to be revisited if and when a switchover to Fedora is contemplated by UTas, and when ADTP accepts links to OAI-PMH compliant repositories.

The ADTP harvesting problem

Clearly it would be counter to its responsibilities to the Australian university community for the University of Tasmania not to be represented in ADTP, and for its research theses not to be accessible to searchers who use this facility. How then can the UTas unified OAI-PMH-compliant repository holding many forms of research output talk to a program with a restricted domain of documents and a much older and non-standard form of harvesting? The

problem is compounded because ADTP will index only theses accepted for research higher degrees, defined as research-only Master degrees and PhDs. Theses for other degrees are not acceptable to ADTP, yet also exist in the UTas repository.

The ADT program is in the process of transitioning to OAI-PMH compliance, but it is not yet clear when this will occur. In the meantime, the harvesting of metadata from all local repositories is done though a module ('Gatherer') in the HotMeta software presently used by ADTP. In brief this looks for an HTML page for each thesis and harvests Dublin Core metadata from the HTML metatags in the pages. A database refresh is achieved by fully reharvesting the metadata.

Overall design

Design principles

In traditional paper-based university libraries, higher degree theses and research articles are viewed and treated very differently. Theses are collected as bound volumes and indexed in the catalogue, usually as a defined physically and logically separate collection. Research articles are not collected, and the authors dispose of their copyright to publishers who make money out of publishing the articles; some of which the library may re-acquire in its serial subscriptions.

As libraries make the transition to digital world, the first of these to come under examination is the digitization of the thesis collection. Many universities around the world have a digital thesis collection, or participate in a union collection of theses. These collections are frequently in a discrete repository. Research articles usually come later as they are not traditional library materials, stimulated by the Open Access movement. The usually voluntary nature of participation in a research article archive as opposed to the compulsory requirements of submitting a thesis for examination, imposes another difference.

Yet viewed objectively, there is absolutely no reason for treating these two forms of research output as different. Both will exist as digital objects. A searcher is likely to be indifferent to whether data retrieved is from a thesis or a research article. Indeed the power of the Internet derives from the synergy of unifying disparate networks and collections.

There are also considerable managerial advantages in using a single digital library for maintaining all forms of research output. A single piece of software is easier to maintain than two; long-term archiving is easier to arrange, registration of the digital library with related service providers (harvesters, search engines, citation services, etc) needs to be done only once, and training of staff is simplified. In addition, the programming of value-adding services to a single piece of software is easier to justify.

The University of Tasmania therefore decided to operate a single unified digital research repository, holding all research output. The repository may even be extended to a statewide repository for *all* Tasmanian research. Even

though at a national level the legacy split between research articles and theses persists, this is seen as transitory and convergence as inevitable. This principle drives subsequent decisions. The key problem is to support discovery through different virtual entry views.

Selection of documents

The documents in the UTas EPrints repository may be any of a number of formats, one of which is 'thesis'. ADTP indexes only *theses for research higher degrees* such as PhD and research Master degrees. The UTas repository also contains in its thesis format theses from honours degrees, coursework master degrees and professional doctorates. The repository serves as a digital archive for these documents as well as a public research output library. For example, only First Class Honours theses are made accessible publicly but all Honours theses are available on-campus to staff.

To comply with the ADTP restrictions, a subset of EPrints documents as defined that:

- have *document format* 'thesis', **and**
- contain the strings 'PhD' or 'research Master' as substrings in their *thesis type* field, using case-independent matching. A modification was made to EPrints data entry so that the thesis type is chosen from a pull-down menu rather than being entered as free text, to ensure uniformity in use of this field. The options in the menu are UNSPECIFIED, Honours, Coursework Master, Research Master, PhD, and Other Degree.

Another option considered was to modify the EPrints formats to delete the thesis format and create two new formats: 'research-only thesis', and 'coursework degree thesis'. This was rejected as not being user-friendly, and not conforming to the probable mind-set of those entering the data.

Harvesting cycle concept

Staff of the School of Computing at UTas wrote new Perl software (HarvestToADT) to scan all documents in the EPrints archive. A special metadata file in a special directory /ADT/ is written for each document meeting the above criterion. The details of the file are elaborated subsequently as required for ADTP's Gatherer module. The harvesting structure and the intermediate directory are shown in Figure 2.

The HarvestToADT.pl script is set up as a cron job in the UTas server to run at regular intervals, currently 19:00 daily. As the archive grows and the run time also grows, this may be made less frequent.

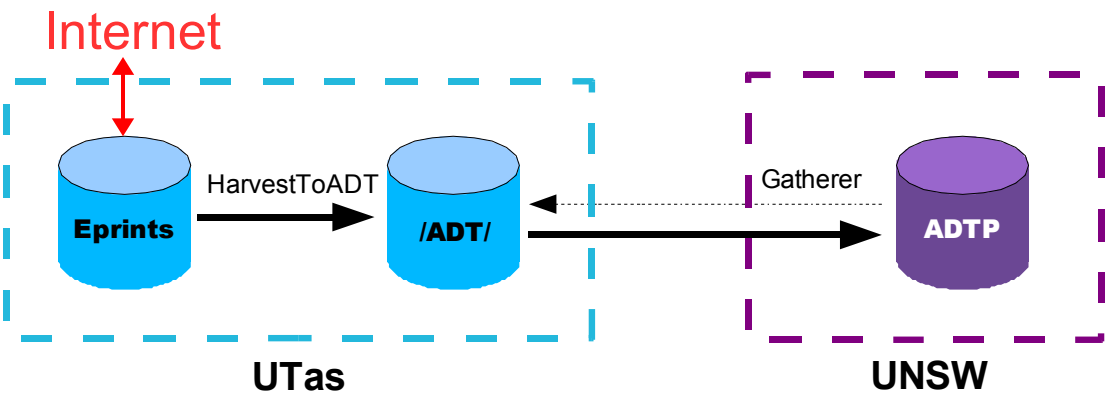

Figure 2 – Harvesting UTas Eprints to ADTP

Detail interfacing

ADTP harvesting cycle

The ADTP harvesting is done by a module in the HotMeta search engine, called *Gatherer*. Once a week, the ADTP database is cleared, and the Gatherer visits every registered local repository, harvesting all records and rebuilding the ADTP metadata database. Once a day, the Gatherer visits each local repository and harvests the metadata from any new document to add to the database.

Directory structure

Each local repository nominates a directory to which the Gatherer is initially pointed. In UTas's case this was chosen to be http://eprints.comp.utas.edu.au:81/ADT/. The Gatherer does an http call on this directory and expects to retrieve an index.html file containing hyperlinks, each to an html file containing metadata for a thesis. The Gatherer then follows all links from this page, retrieving metadata from any pages it finds. The UTas software creates such an index and files.

UTas decided to mimic the ADTP local repository software as closely as possible, so (a) a top level index file in generated, and (b) the /ADT/ directory contains directories, each of which contains the special metadata file. The sub-directories are named according to the ADTP naming scheme; for UTas 'adt-TU1996.0037' where 'TU' is a national library code for UTas, '1996' is the date of the thesis acceptance, and '37' is the EPrints document ID. The actual metadata file is of course titled 'index.html' and is the target of the link in the main index file. See the following indented structure diagram (directories in green).

Structure of /ADT/ directory

```
/ADT/
    index.html
    adt-TU1996.0027/
        index.html
    adt-TU2001.0029/
        index.html
    adt-TU2002.0037/
        index.html
    …
```

File structure

The Gatherer follows any hyperlink it finds, and it is important that it be given no opportunity to escape from the /ADT/ directory. The metadata files are therefore written so as to contain a small amount of html (to save storage space and processing time) and to have absolutely no hyperlinks. The structure chosen by UTas is:

- In the <head> the ADTP metadata as described subsequently.
- In the <body> just a visible title, a visible citation, and a visible EPrints document ID. These allow basic identification of the document for test purposes, but will not be viewed by anyone other than system staff.

To view the actual generated html, point a browser to http://eprints.comp.utas.edu.au:81/ADT/ (view source to see the top-level index page), select one of the links and view source to see that page.

Metadata

The Gatherer extracts the metadata in each html file for the ADTP database. The metadata are generally as described on the ADTP website (ADTP, 2004b). Appendix 1 describes how the metadata are constructed from EPrints fields.

Of particular importance is the URI metadata. This is the full path to the EPrints document metadata (abstract) page, for example http://eprints.comp.utas.edu.au:81/archive/00000037/. A searcher in ADTP will then be sent to the EPrints page when following up the full text of the document. The EPrints page will then display the links to all the full-text files and any ancillary files associated with the thesis. Nothing except the ADTP Gatherer will normally access the /ADT/ directory or files.

EPrints harvesting

The EPrints harvesting by HarvestToADT runs at regular intervals under the control of a cron job. The mySQL databases are interrogated for each EPrints document. If the document meets the selection criterion, then the /ADT/ directory is checked to see if the document is already present. If it is, no further action is taken; if not then a new sub-directory and metadata file are created/written.

EPrints full-text

UTas policy is that the text of a thesis is uploaded to EPrints as a single pdf file. However if desired the text of a thesis can be uploaded as multiple pdf files, for example Contents, Chapter 1, Chapter 2, etc. This option is likely to

be of interest only for scanned theses, which may be large and involve long download times. In normal circumstances, most viewers prefer to see the thesis as a single textual entity. A file naming convention is irrelevant as it is not seen by a viewer, and is not required.

If ancillary materials accompany a thesis such as computer program files, audio or video clips or recordings, animations, slide presentations, etc, these may be uploaded to EPrints in their native format with the thesis text. EPrints allows any kinds or numbers of files to be added to the basic full-text file(s). It is also possible to upload alternate versions of the full-text to assist viewers, for example postscript (.ps), XML (.xml), MS Word™ (.doc), or LaTeX (.dvi).

Linkage to ARROW

The ARROW Discovery Service will eventually hold metadata for all Australia's research output, and because it is OAI-PMH compliant will be harvested by all relevant global search engines. Since ARROW harvests from the UTas unified local repository, and will harvest from ADTP when it becomes OAI-PMH compliant, UTas theses will then be represented twice in ARROW. It has been suggested to ARROW that this is not a cause for concern since the problem is not one created by the UTas repository, but by the present artificial separation between research theses and other research output in the ARROW/ADTP structure. In the long term, ARROW and ADTP will have to look at convergence of their activities.

However should this be seen as a problem, it is not difficult to arrange for research theses and non-theses in the UTas directory to be defined as 'sets', and for the ARROW harvester to be programmed to harvest the non-thesis set alone.

Summary

The decision to establish a unified digital library for the University's research output has been vindicated. The repository is easy to manage and has many opportunities for future development. On the other hand the provision of de-unified set of entry points and harvesting protocols has proved easy to achieve. In effect, the harvester sees a virtual digital library tailored to their collection policy.

The software and assistance with its installation is available to any university which wishes to emulate this implementation, especially in Australia, New Zealand, and neighbouring countries that may wish to link into ADTP. The ADTP linkage software and a few minor modifications to the Eprints upload interface involved a moderate amount of programming, estimated at 30 hours of programming (20 hours development and 10 hours debugging of interface issues).

The University of Tasmania has also commenced to collect an integrated archive, which will serve as a springboard into the emerging world of ever-

more integrated digital libraries, using international standards. A migration path is available to more sophisticated digital repository software.

About the Author

Arthur Sale is currently Professor of Computing Research at the University of Tasmania, and Research Coordinator of its School of Computing. From 1993-99 he was a member of the University's Senior Executive as Pro Vice-Chancellor, and from 1974-93 Chair of the Department of Computer Science. Arthur Sale has published extensively in the ICT literature, and is internationally known for his work on programming languages and computer architecture. His current research interests extend to bioinformatics, health informatics and smart Internet technologies. He lives in beautiful Tasmania, on the shore of the Derwent River.

Acknowledgments

Thanks are due to Christian McGee (University of Tasmania) who performed much of the programming to make this project work, Debbie Campbell (National Library of Australia) for comment, and Fred Piper (University of New South Wales) for assistance with the ADTP interfacing.

References

ADTP (2004a), *Australian Digital Theses Program*, at http://adt.caul.edu.au/

ADTP (2004b), *Metadata Standards*, at http://www.library.unsw.edu.au/thesis/adt-ADT/info/metadata.html, accessed 5 October 2004.

ARC (2004), Scholarly search engine, Old Dominion University, at http://arc.cs.odu.edu/

ARROW (2004), *Australian Research Repositories Open to the World*, at http://arrow.edu.au/

Fedora (2004), *Flexible Extensible Digital Object and Repository Architecture*, at http://www.fedora.info/

EPrints (2004), University of Southampton, at http://eprints.org/

OAI-PMH (2004), *Open Access Initiative Protocol for Metadata Harvesting*, at http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

OAIster (2004), Scholarly search engine, University of Michigan, at http://www.oaister.org/o/oaister/

UTasER (2004), Digital research library, University of Tasmania, at http://eprints.comp.utas.edu.au/

VTLS (2004), VITAL software, at http://www.vtls.com/Products/vital.shtml

Appendix 1 – ADTP metadata

1. **Title**
   <meta name="DC.title" content="Recognition of Sign Language Using Neural Networks">
   Taken from the EPrints title field.

2. **Author**
   <meta name="DC.creator" content="Vamplew, Peter">
   Constructed by concatenating two EPrints author fields in family+first name order. A thesis only ever has one author, so there is no need to look for others in EPrints.

3. **Keywords and phrases**
   <meta name="DC.subject" content="sign language recognition">
   <meta name="DC.subject" content="gesture recognition">
   …
   The EPrints subject field is retrieved. If it contains no commas, the text is split into words and one metadata entry is generated per word. If it contains commas, the text is split instead into phrases based on the comma delimiters.

4. **Abstract**
   <meta name="DC.description"
   content="This thesis details the development of a computer system (labelled the SLARTI system) capable of recognising a subset of signs from Auslan (the sign language of the Australian Deaf community), based on the pattern classification paradigm of artificial neural networks…">
   Taken from the EPrints abstract field.

5. **Date thesis accepted for degree**
   <meta name="DC.date" scheme="W3CDTF" content="1996">
   Taken from the EPrints date field.

6. **Language**
   <meta name="DC.language" scheme="RFC3066" content="en">
   Fixed metadata at present in UTas, as only English language theses are accepted by UTas.

7. **Institution/School**
   <meta name="DC.publisher" content="University of Tasmania, School of Computing">
   Created by concatenating two EPrints fields: Institution and Department.

8. **Copyright**
   <meta name="DC.rights"
   content="http://www.utas.edu.au/copyright/copyright_disclaimers.html”>
   <meta name="DC.rights" content="(c) Copyright 1996 Peter Vamplew">
   The first is the institution-wide disclaimer and is automatically generated in the metadata for all theses; the second is the author’s copyright notice, constructed from the EPrints date and author fields (the name now being in first+family name order).

9. **Universal Resource Identifier**
<meta name="DC.identifier"
content="http://eprints.comp.utas.edu.au:81/archive/00000037/">
This is the URL pointing to a public view of the thesis. In UTas this is the full path to the EPrints view of the thesis metadata.